TECHNOLOGY

MAY 2017

Big Data Guidelines

Information and Privacy
Commissioner of Ontario

Commissaire à l'information et à la
protection de la vie privée de l'Ontario

ACKNOWLEDGEMENTS

We would like to thank the following individuals and organizations for reviewing an earlier version of these guidelines and providing helpful comments:

- Dr. Khaled El Emam
- Information, Privacy and Archives Division, Ministry of Government and Consumer Services of Ontario
- Office of the Information and Privacy Commissioner of British Columbia

CONTENTS

INTRODUCTION

As governments strive to develop more informed policies and procedures about issues affecting the public, they sometimes seek to gain new insights and obtain additional evidence through the use of combined data sets of linked information about individuals. To create and analyze these large data sets, government institutions are increasingly turning to the use of "big data."

The term "big data" generally refers to the combined use of a number of advancements in computing and technology, including:

- *New sources of personal information*. The continuing digitization of records and services, as well as the widespread use of smart devices and social media, has dramatically increased the amount and kinds of information that are available about individuals.
- *Virtually unlimited capacity to store data*. Significant reductions in the cost and physical size of devices, combined with parallel increases in storage capacity, have resulted in a situation where vast amounts of data can be stored for little cost.
- *Improved record linkage techniques*. Techniques to link data together include both "deterministic" approaches, where records are matched only if they have the same unique identifier, and "probabilistic" or "fuzzy" approaches, which allow for variations in the values of identifiers and match records based on calculated probabilities.
- *Computer programs that can learn from and make predictions on data*. Increases in computing power as well as advancements in statistical and mathematical procedures have led to the creation of algorithms which can analyze large, complex data sets and uncover hidden patterns and correlations in the data to derive rules or insights, which in turn can be used to explain phenomena or build predictive models that allow for automated decision-making.

Big data can be an important tool for shaping and improving government policies, programs and services. For example, public health and the provision of health care may be improved by using big data to analyze disease patterns and outbreaks to discover unknown sources or contributing factors. Another potential benefit is the detection of fraud. Suspicious patterns of activity can be detected by big data and used to determine if there are reasonable grounds to investigate further.

While big data may provide benefits to individuals, it also raises a number of privacy, fairness and ethical concerns with respect to the processing of personal information. Government institutions with the legal authority to use big data should understand and address these concerns in order to prevent uses of personal information that may be unexpected, invasive, inaccurate, discriminatory or disrespectful, as discussed in these guidelines. The purpose of these guidelines is to inform such institutions of the key issues to consider and best practices to follow when conducting big data projects involving personal information.

These guidelines are not a comprehensive assessment of every issue or best practice involving big data projects in which personal information may be collected, used or disclosed. When conducting a big data project, it is important that you consider all applicable legislation, including Ontario's privacy laws and their regulations, and seek advice from your freedom of information and privacy coordinator or legal counsel, where appropriate.

SCOPE OF GUIDELINES

The use of big data by government institutions is an important and timely topic, but also a complex and challenging one, engaging new and emergent information technologies that may be used in various ways to fulfill a wide range of policy goals. The Office of the Information and Privacy Commissioner of Ontario (IPC) will continue to work on issues related to big data and plans to release additional guidance documents aimed at specific sectors of government and at providing further information on some of the best practices identified in this document.

Another challenge with big data is that its technical nature can make it difficult to provide guidance to a non-specialist audience. These guidelines attempt to strike an appropriate balance between presenting issues at too high a level so as to offer little substantive discussion or practical guidance and presenting them at too low a level so as to become immersed in details and complexities. The goal is a document that is both accessible and useful to institution program managers, freedom of information and privacy coordinators and technical staff. For readers who would like a basic introduction to big data, please see the IPC's Fact Sheet "Big Data and Your Privacy Rights."[1] For readers interested in exploring the issues discussed in this document in further detail, a list of resources that may be helpful is provided in Appendix A.

[1] Office of the Information and Privacy Commissioner of Ontario, "Big Data and Your Privacy Rights," January 2017, https://www.ipc.on.ca/wp-content/uploads/2017/01/fact-sheet-big-data-with-links.pdf.

BIG DATA AND ONTARIO'S PRIVACY LAWS

As it currently stands, many of the information practices involved in big data would not be compliant with the privacy protections set out in Ontario's public-sector privacy laws, the *Freedom of Information and Protection of Privacy Act (FIPPA)* and the *Municipal Freedom of Information and Protection of Privacy Act (MFIPPA)*. Not only were *FIPPA* and *MFIPPA* not designed with big data-type practices in mind, the practices themselves were not even possible at the time. When *FIPPA* came into effect in 1988, it was not yet possible to browse a website or receive services online, because the World Wide Web had not yet been invented. When *MFIPPA* came into effect three years later, commercial Internet service providers were only beginning to emerge. The use and availability of information technology was nowhere close to the levels we see today. If personal information was needed, it was typically for discrete purposes that were determined in advance. Complex data types and advanced analytics were not yet a reality. The current legislative framework is based on a set of protections that in effect require government institutions to act as "silos" of personal information. The protections in *FIPPA* and *MFIPPA* include:

- the collection of personal information must be "necessary"
- secondary uses are restricted
- information sharing is limited

Despite the above, it may still be possible to conduct big data projects within the context of *FIPPA* and *MFIPPA* in cases where the required practices are authorized under another law—for example, in the enabling legislation of a government ministry. *FIPPA* and *MFIPPA* may allow for big data-type practices if the collection of personal information is "expressly authorized by statute"[2] and any disclosures are for the purpose of "complying with an Act of the Legislature."[3] The best practices developed in these guidelines are intended to assist government institutions which have the legal authority to conduct big data projects.

Although big data projects may be possible within the context of *FIPPA* and *MFIPPA*, such instances should be the exception, not the rule. To allow for big data-type practices in general, a new or modified legislative framework is needed. In addition to providing guidance to institutions with the authority to conduct big data projects, the best practices developed in these guidelines can also be viewed as a non-exhaustive list of recommended elements of a regulatory and policy framework to enable big data projects while protecting the privacy of individuals and ensuring the fair and ethical use of their personal information.

[2] See section 38(2) of *FIPPA* and section 28(2) of *MFIPPA*.

[3] See section 42(1)(e) of *FIPPA* and section 32(e) of *MFIPPA*.

Note that any framework that enables big data projects should include provisions to ensure effective and independent oversight and require appropriate notification in the event of a breach of personal information or violation of individual rights. These and other strongly recommended elements of a regulatory and policy framework are not discussed in these guidelines.

KEY CONSIDERATIONS AT EACH STAGE OF A BIG DATA PROJECT

A big data project can be a complex undertaking involving multiple stages, including defining the business use case, planning and conceptualization, getting support from senior management and building a team with the required expertise. This is in addition to stages that directly involve the collection, use and disclosure of personal information. To help focus the discussion of issues and best practices in this document, the process of conducting a big data project has been divided into four stages that involve the processing of personal information in some form:

1. collection
2. integration
3. analysis
4. profiling

Not every big data project will involve all four of these stages. Projects that collect data sets from secondary sources and integrate them, but whose analysis only quantifies statistical properties or explains patterns in the data and do not build predictive models or profiles will not involve the fourth stage, “profiling.” Other projects may only involve the third stage, “analysis,” if no data sharing, indirect collection or profiling is involved.

STAGE 1: COLLECTION

The first stage of a big data project consists in the identification and collection of multiple data sets from various sources of personal information. Each data set will likely contain a different combination of data points or values about the individuals whose personal information is being collected.

When collecting personal information as part of a big data project, you should consider the impact of a number of issues, including:

- indirect collection and secondary purposes
- speculation of need rather than necessity
- public notification
- privacy of publicly available information

INDIRECT COLLECTION AND SECONDARY PURPOSES

At the heart of big data lies a fundamental tension with some basic tenets of privacy and the protection of personal information. Many, if not all, of the issues that arise are the result of big data's incompatibility with two of the most fundamental principles of data protection—that (i) personal information should be collected directly from the individual to whom it pertains, and (ii) it should only be used for the purpose for which it was collected (with limited exceptions). Big data promotes neither of these principles. In general, big data involves information that has been collected indirectly, and used for a purpose which may not have been intended at the time of the original collection.

Although this tension is real and pressing, it is not irresolvable. With the appropriate safeguards in place, it is possible to protect the privacy of individuals and ensure the fair and ethical processing of their personal information while conducting big data projects. While specific safeguards are discussed later in this document, a big data project should have the legal authority to directly or indirectly collect any personal information involved in it and use the information for the purposes of the project.

> **Best practice:** Ensure that you have the legal authority to directly or indirectly collect any personal information involved in your big data project and use it for the purposes of the project.

SPECULATION OF NEED RATHER THAN NECESSITY

The collection practices of big data are informed by the unique approach it takes to the analysis of information. Big data does not start out with a preconceived rule or hypothesis and then look to the data as a means of supporting or proving it, as was common in traditional, "little" data analyses. Instead, big data "fishes" for statistically significant patterns or correlations without prior knowledge of what they are and, in certain cases, why they may be useful.

An issue that emerges from this type of analysis is that big data projects often find themselves at odds with another fundamental principle of data protection—data minimization or the practice of limiting the collection of personal information to that which is directly relevant and necessary to achieving a specified purpose. If the rule or hypothesis to be derived is not known in advance of analyzing the information, how can you select a minimal set of data elements to support or prove it? In other words, how can a set of data elements be "directly relevant" and "necessary" when their respective utility or role in the overall analysis may not be known at the time of collection?

Although all inquiry, by nature, presupposes some lack of knowledge in the underlying subject-matter, no big data project, like any scientific endeavour, should base its collection of data elements on mere speculation. A common refrain in data science is: "Garbage in, garbage out!" The more relevant the data, the greater the chances of success. Although the rule or hypothesis may not be known in advance of the analysis when using big data, at a minimum, the collection of data elements should be conceptually related to the subject-matter under investigation and be directly informed by the question being asked. Moreover, the purpose of a big data project should always be tied to the mandate of the institution.

To protect privacy and ensure the fair and ethical processing of personal information, a collection of personal information done within the context of a big data project should be reviewed and approved by a research ethics board (REB) or similar body. The REB reviewing the collection practices of the big data project should consider a number of factors when deciding whether to approve a big data project, including:

- whether the personal information that is to be collected is reasonably limited, taking into consideration the objectives of the project and the nature of the mathematical and statistical procedure to be used
- whether the potential benefits to be derived from the project outweigh the foreseeable risks to the individuals whose personal information is being collected
- whether adequate safeguards will be in place to protect the privacy of the individuals whose personal information is being collected and to preserve the confidentiality of the information
- the potential for the personal information that is to be collected to stigmatize, discriminate or otherwise result in the unfair treatment or consideration of an individual or group of individuals

In addition to collection practices, a REB should also consider the privacy, fairness and ethical implications of the integration, analysis and profiling stages of a big data project, if applicable, in its review and approval of the project.

The level of review required of a big data project can vary depending on the level of risk it presents to individuals and groups. Big data projects that present lower levels of risk can receive less scrutiny than projects presenting higher levels of risk. The REB should be comprised of individuals with the necessary knowledge, expertise or representation in areas relevant to the project, such as:

- research ethics
- data science and analytics
- privacy and other relevant laws
- the public or community membership

Best practice: Ensure that the privacy, fairness and ethical implications of your big data project are reviewed and approved by a research ethics board or similar body.

PUBLIC NOTIFICATION

The indirect collection and secondary use of personal information at the heart of big data creates additional challenges to the openness and transparency of big data projects. The nature of big data makes it difficult, if not impossible, to notify individuals at the time of the original collection in any meaningful way about the existence or purpose of big data projects involving their personal information. However, if individuals are to have a say in how their personal information is processed, they must be aware, or have a means of becoming aware, of the full extent to which it is collected, used and disclosed. How can individuals become aware of big data projects involving their personal information?

To promote openness and transparency, a description of each big data project should be published on the host institution's website to enable individuals to become informed about how their personal information is being processed. The description should contain relevant information about the big data project, including:

- the title of the project
- the purpose and public benefit of the project
- a description of the data sets involved, including their sources, and the procedure used to analyze the data
- the output of the analysis
- retention schedules for the data sets involved

Best practice: Ensure that you publish a description of your big data project on your institution's website.

PRIVACY OF PUBLICLY AVAILABLE INFORMATION

While in the past it may have seemed appropriate to assume that individuals forfeit any right to privacy in personal information about themselves they make available online, in the context of big data, this position is increasingly problematic. When using big data, the potential uses and insights that can be derived from a piece of information are no longer discrete and recognizable in advance. Personal information that may be innocuous on its own can be collected, integrated and analyzed with other sets of personal information to reveal hidden patterns and correlations that only an advanced algorithm can uncover due to the size and complexity of the information. Because individuals would likely not expect their personal information to be used in such ways, institutions should assume that individuals have a reasonable expectation of privacy in publicly available information.

To protect the privacy of individuals, you should consider treating personal information that is publicly available the same as non-public personal information when conducting big data projects.

Best practice: Consider treating any publicly available personal information involved in your big data project the same as non-public personal information.

STAGE 2: INTEGRATION

Once you have identified and collected the data sets involved in your big data project, the second stage consists in combining and linking the information together to form a single integrated data set. Also known as "compilation" or "consolidation," this stage is primarily concerned with preparing the information for analysis.

The topic of data integration—sometimes called "data linking/linkage" and "data/computer matching"—predates big data and concerns regarding it have been raised since at least the creation of data protection and privacy laws. For example, the 1980 report of the Ontario Williams Commission (*Public Government for Private People: The Report of the Commission on Freedom of Information and Individual Privacy*) provides a high-level description of some of the issues that arise from combining and linking data sets together:

> The prospects of greater integration of data bases raises, in turn, a number of informational privacy issues [...]. The possibility that information gathered for one purpose might be used for quite a different purpose is enhanced. The use of data linkage may increase the likelihood that decisions will be based on erroneous information, or on the basis of an individual's historical record rather than his current circumstances or more recent pattern of conduct. In short, it is feared that the use of such dossiers may constitute a form of data surveillance which might operate against the legitimate interests of the individual.[4]

Not only do the above issues continue to be relevant today, but their importance is only magnified within the context of big data. Although they remain valid, some of these issues are discussed elsewhere in this

[4] *Public Government for Private People: The Report of the Commission on Freedom of Information and Individual Privacy*, vol. 3 (Toronto: Queen's Printer, 1980), 771.

document. In this section, the issues under consideration are limited to those that arise from the act of combining and linking data sets together.

When integrating data sets containing personal information as part of a big data project, you should consider the impact of a number of issues, including:

- linking errors from probabilistic linkages
- inadequate separation of policy research and administrative functions
- creation of new databases

LINKING ERRORS FROM PROBABILISTIC LINKAGES

For data sets to be linked together, they must share a unique identifier or group of identifying fields about the individuals whose personal information is contained in them. Only data sets that share a unique identifier or group of identifying fields can compare their respective values for matches. Each match constitutes a combination of records that may be linked together on the basis that they refer to the same individual.

Because of the diversity of sources of personal information, it is rare for data sets collected as part of a big data project to share a unique identifier—for example, a health card number—or to have a group of identifying fields with consistently high data quality. However, if this is the case, the linking procedure is straightforward: do a direct comparison of identifiers and link the records together with identical values. This procedure is known as "deterministic" linkage.

What is more likely is that the data sets will share a group of identifying fields but have inconsistencies in their values that may be caused by differences in data quality and formatting. For example, in a group of fields containing first name, middle name, last name, date of birth and gender, the same individual's first name may be spelled "Michael" in one data set but "Mike" in another; one data set may contain the full middle name whereas another only has the first initial; and the date of birth may be recorded as "February 11, 1991" in one data set but as "February 12, 1991" in another.

To account for such variations, non-deterministic linking procedures typically work by calculating the probability that two records refer to the same individual and then comparing that probability to two thresholds: a "match" and a "non-match" threshold. If the probability is lower than the non-match threshold, the records are not considered a match. If the probability equals or exceeds the match threshold, the records are considered a match and linked together. Probabilities that fall in between the two thresholds are reviewed manually. This procedure is known as "probabilistic" or "fuzzy" linkage.

The non-deterministic nature of probabilistic linkages as well as the selection of identifying fields in both deterministic and probabilistic linkages raises issues related to another fundamental principle of data protection—the data quality principle or the principle that personal information should be accurate, complete and kept up-to-date to the extent necessary to fulfill the purposes of its use. If a comparison of records is probabilistic, involving a degree of variability in the values of identifying fields, what is an appropriate threshold to determine the existence of a match or non-match? Indeed, if the data sets to be linked do not share a unique identifier, what is an appropriate group of fields to identify individuals uniquely across the data sets?

While the data quality principle recognizes the importance of maintaining the accuracy of personal information, the level of accuracy it requires is not absolute, but rather depends on the proposed use of the information. Applied to record linkages, this means that the required level of accuracy of a linking procedure may vary depending on the use of the linked data sets and the purposes of the big data project. For example, linked data sets used for the purposes of drawing conclusions about a population as a whole would generally be held to a lower standard of accuracy than linked data sets used for the purposes of making decisions about specific individuals.

When integrating data sets as part of a big data project, you should ensure that the record linkage procedure is accurate to the extent necessary to fulfill the purposes of the project.

> **Best practice**: Ensure that your record linkage procedure is accurate to the extent necessary to fulfill the purposes of your big data project.

INADEQUATE SEPARATION OF POLICY ANALYSIS AND ADMINISTRATIVE FUNCTIONS

The overall delivery of a government program (or service) can be divided into two basic functions:

1. an administrative function in which the program is delivered directly to members of the public

2. a policy analysis function in which the program undergoes planning and evaluation, including policy development, system planning, resource allocation and performance monitoring

In most cases, personal information collected for the purpose of administering a program can be used for the secondary purpose of fulfilling the policy analysis function of the program. If individuals have participated in a government program, it is generally considered a “consistent purpose” for their personal information to be used subsequently in the planning and evaluation of the program.

However, what may be considered a “consistent purpose” does not work in the opposite direction. Personal information collected as part of the policy analysis function of a program cannot, in general, be used subsequently in the administration of the program. There are two reasons for this. First, individuals who volunteer their personal information for a policy analysis purpose are often assured, like any primary research project, that their participation will remain confidential, unless the findings of the analysis directly benefit them. In the absence of any direct benefit, the reuse of participants’ personal information to make decisions about them individually would go against this assurance.

Second, policy analysis activities are typically only concerned with individuals insofar as their information is required to draw conclusions about a population as a whole. By making general observations rather than specific decisions about individuals, most policy analysis offers its participants a level of privacy protection by default. To void this protection by reusing participants’ personal information for administrative purposes would compromise the integrity of the project and erode public trust in the process.

When conducting a big data project to fulfill the policy analysis function of a program, government institutions may end up collecting and integrating data sets of personal information that go beyond the personal information collected for the purpose of administering the program. This potential for separate data sets collected for incompatible purposes to arise within the same institution raises the issue of their functional separation. How can the use of personal information to fulfill the policy analysis function of a program be separated from the use of personal information as part of the administration of the program?

Administrative and policy analysis functions require different classes of information to fulfill their respective purposes. While the administration of a program must use identifiable information to deliver the program to specific individuals, policy analysis can use non-identifiable information to draw conclusions about a population as a whole. Because of these differences, administrative and policy analysis functions can be separated through *de-identification*.[5] If linked data sets are de-identified, they can only be used to fulfill the policy analysis function of a program and cannot be used in the administration of the program.

[5] De-identification is the process of removing any information that identifies an individual, or for which there is a reasonable expectation that the information could be used, either alone or with other information, to identify an individual, while preserving as much utility in the information as possible. See Office of the Information and Privacy Commissioner of Ontario, *De-identification Guidelines for Structured Data*, June 2016, https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf.

The de-identification of integrated data sets also acts as an important mitigation measure in helping to address the inherent tension between big data and the principle of data minimization. Although de-identification does not limit the scope of data elements collected and integrated as part of a big data project, it adds a layer of privacy protection after the fact insofar as it reduces the identifiability of the information to be analyzed. A benefit of this is that it helps to protect against theft, loss and unauthorized disclosures of personal information, in addition to unauthorized uses.

When integrating data sets as part of a big data project, you should de-identify any personal information in the linked data sets to ensure adequate separation between your policy analysis and administrative functions. In addition to de-identification, you may also wish to explore the effectiveness of a number of emerging technologies in the area of privacy-protective data analysis.[6]

Best practice: De-identify your linked data sets to ensure adequate separation between your policy analysis and administrative functions.

CREATION OF NEW DATABASES

Data sets that have been collected and integrated as part of a big data project can reveal a great deal about the individuals whose personal information is contained in them. While individual pieces of personal information may not reveal much about an individual on their own, when linked together and compiled from a variety of sources or over time, an increasingly detailed portrait of individuals' lives may begin to emerge. When this happens within the context of a big data project, it can lead to the creation of a new government database containing a disproportionate amount of information about the personal aspects of individuals. The sensitivity and comprehensiveness of such a database would make it an attractive target for unauthorized access, theft or use in ways that disadvantage certain individuals or groups.

While a big data project may require increases in the amount and types of personal information collected and integrated by government institutions, in most cases there is no practical requirement to retain the information indefinitely. To be successful, a big data project need only retain personal information for the duration of the project.

[6] Although not yet fully mature, formal mathematical approaches to privacy, such as differential privacy and synthetic data, as well as advanced cryptographic techniques, such as secure multiparty computation and fully homomorphic encryption, may also protect the privacy and confidentiality of individuals while allowing integrated data sets to be analyzed.

When integrating data sets as part of a big data project, you should ensure that the integration does not result in the creation of a new permanent database of personal information, and that all copies of data sets containing personal information are destroyed as soon as is reasonably possible.

> **Best practice**: Ensure that your integration of data sets does not result in the creation of a permanent database of personal information, and that all copies of data sets containing personal information involved in your big data project are destroyed as soon as is reasonably possible.

STAGE 3: ANALYSIS

Once you have integrated the data sets collected as part of your big data project, the third stage of the process consists in analyzing them to derive new insights and findings. Depending on the objectives of your big data project, the type of mathematical and statistical procedure used to analyze the information may differ. In general, there are three possibilities, each resulting in a different type of outcome. When analyzing information, a big data project may:

1. quantify properties, resulting in summary statistics
2. uncover hidden patterns and correlations, resulting in a rule or explanation of some phenomenon
3. build a predictive model, resulting in a profile of individuals

Although different procedures may be used to analyze information as part of a big data project, the issues that arise at this stage of the process do not stem from the type of procedure used so much as from the selection and composition of the data sets themselves. When analyzing information as part of a big data project, you should consider the impact of a number of issues, including:

- poor data quality
- biased data sets
- discriminatory proxies
- spurious correlations

POOR DATA QUALITY

The greater the variety of data sets analyzed as part of a big data project, the greater the potential for errors and inconsistencies to appear in the information. Data quality issues can be exacerbated by the diversity of sources and types of personal information involved in big data projects.

As discussed above, the data quality principle of data protection provides that the level of quality required of personal information depends on its proposed use. In the case of analysis, this means that the required level of accuracy, completeness and currency of a data set may vary depending on the purpose of the analysis and type of procedure used. For example, if the purpose of the analysis is to gain insight into a complex issue with a high degree of precision, then the quality of the data set will likely have to be higher than in the case of an insight into a general trend.

The size of the data set may also impact the level of data quality. For example, if the purpose of the analysis is to gain insight into a general trend, the quality of a data set with more information may not need to be as high as a data set with less information.

When analyzing an integrated data set as part of a big data project, you should ensure that the data set is accurate, complete and up-to-date to the extent necessary to fulfill the purposes of the project.

Best practice: Ensure that the information analyzed in your data sets is accurate, complete and up-to-date to the extent necessary to fulfill the purposes of your big data project.

BIASED DATA SETS

Big data is sometimes celebrated for the fact that it can analyze “all” the data and does not require the collection of samples, which only represent a subset of the target population. With sampling, care must be taken to ensure that the individuals selected for analysis accurately represent the target population. If the method of collecting samples is not properly randomized, the resulting data set may be “biased” in the sense that it excludes certain members of the population. Without the need for sampling, big data is sometimes characterized as being more objective and unbiased than traditional, “little” data analyses.

Although big data does not require the collection of samples, the data sets collected and integrated as part of a big data project are still susceptible to bias. Sampling is not the only way in which individuals or groups may be selected for inclusion or exclusion in a data set. Even if “all” the data is collected, the practices that generate the data in the first place may contain implicit biases that over- or underrepresent certain members of the population. For example, if the practices of an organization allow for subjective or discretionary decisions to be made about individuals and these decisions disproportionately single out certain individuals over others, then the data representing those practices will simply reflect this imbalance.

Consider the case of hiring decisions. If the hiring practices of an organization have resulted in people from similar backgrounds being hired more often, then the data on those hires will reflect those decisions. If that data is then analyzed to find common attributes to screen future applicants, the biases of the earlier hiring practices can be reinforced.

Bias may also enter into data sets as a result of poor design in the delivery of a program or service. For example, if a program that is open to the public contains technical or socio-economic barriers that prevent certain groups of individuals from participating, then the data representing the outcome of the program will simply reflect this exclusion. Such biases are often the result of overly restrictive program requirements.

When analyzing an integrated data set as part of a big data project, you should ensure that it is representative of the target population to the extent necessary to fulfill the purposes of your big data project. When assessing the representativeness of a data set, you should consider a number of factors, including:

- whether the practices that generated the data set allowed for discretionary decisions
- whether the program or service contained requirements that were overly restrictive

> **Best practice**: Ensure that the information analyzed in your data sets is representative of the target population to the extent necessary to fulfill the purposes of your big data project.

DISCRIMINATORY PROXIES

Section 15 of the *Canadian Charter of Rights and Freedoms* (Charter) guarantees every individual a right to “equal protection and equal benefit of the law without discrimination,” and in particular without discrimination based on “race, national or ethnic origin, colour, religion, sex, age or mental or physical disability.” This right to non-discrimination extends to the collection, use and disclosure of personal information by government institutions.

While the Charter prohibits the unfair treatment or consideration of individuals based on certain protected personal characteristics, the diversity of sources and types of personal information involved in big data projects can create challenges to institutions’ compliance with this requirement.

This is especially true in cases where the analyzed information contains a variable that is not itself explicitly protected but correlates with a protected characteristic. For example, if a geographic region contains a high percentage of individuals with the same racial or ethnic background, then a big data project that analyzes geographic regions to build a profile of the individuals living in them may result in decisions being made about those individuals that are, in effect, based on race and ethnicity.

When analyzing an integrated data set as part of a big data project, you should be aware of the potential for variables to correlate with protected personal characteristics and ensure that your analysis does not result in any such variables being used as proxies for prohibited discrimination. In addition to the information involved, the outcome of the analysis may need to be reviewed by a REB or similar body to determine its potential for such discrimination.

> **Best practice**: Be aware of the potential for variables to correlate with protected personal characteristics and ensure that the analysis of your integrated data set does not result in any such variables being used as proxies for discrimination.

SPURIOUS CORRELATIONS

One of the purposes of analyzing information is to discover patterns or statistical relations which may indicate meaningful relationships among the variables involved. On the basis of such discoveries, insights into the corresponding subject matter may be generated and rules for predictive models may be derived.

While increases in computing power and the development of advanced algorithms have enabled big data to detect the presence of increasingly complex relationships among increasingly large numbers of variables, this ability brings with it an all-important risk. With so many combinations of variables at play, there are likely to be some that appear to be meaningful without actually being so. When analyzing an integrated data set as part of a big data project, understanding the difference between correlation and causation is key.

Correlation means that the values of two variables in a data set are statistically related. For example, they tend to increase or decrease together. Causation is the stronger claim that the variables relate by necessity and that a change in one always brings about a change in the other. Discovering a correlation, however, does not necessarily mean that the change in one variable was the cause of the change in the other variable. The two could simply relate by chance, in which case the relationship between them would be coincidental rather than causal, or they could both be related to a third variable that was not considered. While two variables that share a causal relation should always correlate in a data set, a correlation by itself does not imply causation.

When analyzing an integrated data set as part of a big data project, you should be aware of the potential for spurious correlations and ensure that any patterns discovered in the analysis are meaningful. You may need to verify the results of the analysis in a manner that is independent of the procedure used in order to ensure the meaningfulness of certain patterns.

> **Best practice**: Be aware of the potential for spurious correlations and ensure that any patterns discovered in the analysis of your integrated data set are meaningful.

STAGE 4: PROFILING

Only big data projects that build a predictive model or profile of individuals as a result of the analysis conducted in stage three will involve the fourth stage of the process. This stage consists in using the now built model to evaluate or predict attributes of individuals on a case-by-case basis.

In the context of big data, profiling is a type of automated processing of personal information. It works by taking an individual's personal information and inputting it into a predictive model, which then processes the information according to the set of rules established by the model to produce an evaluation or prediction concerning one or more attributes of the individual.

Depending on the objectives of the big data project, profiling may be used to evaluate or predict different attributes of individuals. For example, it may be used to evaluate or predict an individual's eligibility for programs or services, economic situation, health, behaviour or movements.

When using a predictive model or profile to evaluate or predict attributes of individuals as part of a big data project, you should consider the impact of a number of issues, including:

- lack of transparency
- false predictions
- individuals as objects

LACK OF TRANSPARENCY

It is important to note that profiling does not only process personal information but generates it as well. The evaluation or prediction of an individual's personal attributes results in the creation of a new element of personal information that will be associated with the individual.

This aspect of profiling raises issues of transparency. While individuals should be aware of any personal information that is collected directly from them, the generation of personal information as a result of profiling happens in the background and is less conspicuous. An individual who is the subject of profiling may not be aware of the fact that in addition to the elements of personal information collected directly from them, profiling has generated additional fields of personal information.

The inconspicuous nature of profiling may also lead to individuals not understanding the consequences it may have on them and to a lack of transparency around the decision-making process. If not properly designed and implemented, profiling may result in significant decisions being made about individuals without their knowledge based on information that they may not have wanted to share or felt comfortable sharing, leading to unexpected results.

To promote transparency, individuals who are the subject of profiling should be informed of additional information regarding the nature of the predictive model or profile being used, including:

- the use of profiling and the fields of personal information generated by it
- a plain-language description of the logic employed by the predictive model
- the implications or potential consequences of the profiling on individuals

> **Best practice**: Ensure that individuals who are the subject of profiling are informed of additional information regarding the nature of the predictive model or profile being used.

FALSE PREDICTIONS

A model is only ever a snapshot of the reality it aims to represent. Although predictive models may strive for perfection in terms of their accuracy, not only is this difficult to achieve in practice, but it is debatable if such a distinction is even achievable at all, especially when the prediction concerns the behaviour of human beings, whose practices, values and goals are constantly evolving. No matter how much data or how many data points go into the calculations of a predictive model, some degree of error is to be expected.

Profiling may be used in different ways to make decisions of varying degrees of significance about individuals. For example, a big data project may make decisions about individuals based solely on the results of profiling or a project may use profiling to provide a (human) decision-maker with an additional factor to consider in a multi-factor decision.

In cases where profiling is used as the sole basis for a decision that significantly affects an individual, a false prediction may not only result in the individual being improperly treated, but significantly harmed as well. For example, if a decision that leads to a denial, termination, suspension or reduction of a benefit or entitlement is based solely on the results of profiling, a false prediction would not only improperly treat but significantly harm the affected individual.

Another issue to consider is who should be responsible for ensuring the accuracy of such decisions. It would not be fair to deny or reduce a benefit or entitlement to an individual based solely on the results of profiling and then place the burden on the individual to correct any errors. It is also important to note that some vulnerable individuals may be limited in terms of their ability to navigate the system of administrative procedures required to challenge an incorrect finding.

When using profiling as part of a big data project, you should verify the results of any decisions based solely on profiling in cases where the decisions significantly affect individuals and ensure that individuals are given the opportunity and sufficient support to challenge or respond to such decisions. The results should be verified in a manner that is independent of the predictive model or profile used.

Best practice: Verify the results of decisions based solely on profiling in cases where the decisions significantly affect individuals and ensure that individuals are given the opportunity and sufficient support to challenge or respond to such decisions.

INDIVIDUALS AS OBJECTS

Profiling is made possible through the practice of placing individuals into predefined types or categories. Only on the basis of such a reductive approach to understanding individuals can a decision-making process be automated. Individuals only amount to the sum of their parts when they are profiled.

This aspect of profiling raises ethical issues involving the effects of profiling and automated decision-making on individuals and society. While such issues go beyond the traditional notion of privacy, they nonetheless engage concepts and ideas that form the basis for why privacy is important and a right valued by Ontarians.

Even if the use of profiling is transparent and produces accurate predictions, individuals may still feel a loss of dignity or respect as a result of their being subjected to profiling. By its very nature, profiling treats individuals as fixed, transparent objects rather than as dynamic, emergent subjects.

In addition to a loss of dignity or respect, profiling may have larger effects on society and individuals. Assume for the moment a predictive model with perfect accuracy. While use of such a model would obviously result in increases to the efficiency of programs and services, it is also clear that the extension of such a model to too many aspects of society or individuals' lives would have serious consequences. Individuals would gradually lose or have no use for their autonomy. Chance occurrences and fortunate discoveries may cease to happen. Individuals would no longer be exposed to a variety of perspectives and different opinions.

When using profiling as part of a big data project, you should consult with the public and civil society organizations regarding the appropriateness and impact of the proposed use of profiling and provide them with an opportunity to comment on the effects the profiling may have on society and individuals' lives.

> **Best practice**: Consult with the public and civil society organizations regarding the appropriateness and impact of any proposed use of profiling and provide them with an opportunity to comment on the effects the profiling may have on society and individuals' lives.

SUMMARY OF BEST PRACTICES

While big data can be an important tool for shaping government policies, programs and services, it raises a number of privacy, fairness and ethical concerns that need to be addressed by government institutions in order to prevent uses of personal information that may be unexpected, invasive, inaccurate, discriminatory or disrespectful of individuals. To address these issues, institutions with the authority to conduct big data projects should follow the set of best practices developed in these guidelines. These best practices apply to a four stage process for conducting big data projects.

During the **collection stage** of a big data project, institutions should:

- Ensure that they have the legal authority to directly or indirectly collect any personal information and use it for the purposes of the project.
- Ensure that the privacy, fairness and ethical implications of the project are reviewed and approved by a research ethics board or similar body.

- Ensure that they publish a description of the project on their website.
- Consider treating any publicly available personal information involved in the project the same as non-public personal information.

During the **integration stage** of a big data project, institutions should:

- Ensure that the record linkage procedure used is accurate to the extent necessary to fulfill the purposes of the project.
- De-identify the linked data sets to ensure adequate separation between their policy analysis and administrative functions.
- Ensure that the integration of data sets does not result in the creation of a permanent database of personal information and that all copies of data sets containing personal information involved in the project are destroyed as soon as is reasonably possible.

During the **analysis stage** of a big data project, institutions should:

- Ensure that the information analyzed is accurate, complete and up to date to the extent necessary to fulfill the purposes of the project.
- Ensure that the information analyzed is representative of the target population to the extent necessary to fulfill the purposes of the project.
- Be aware of the potential for variables to correlate with protected personal characteristics and ensure that the information analyzed does not result in any such variables being used as proxies for discrimination.
- Be aware of the potential for spurious correlations and ensure that any patterns discovered in the analysis are meaningful.

During the **profiling stage** of a big data project, institutions should:

- Ensure that individuals who are the subject of profiling are informed of additional information regarding the nature of the predictive model or profile being used.
- Verify the results of decisions based solely on profiling in cases where the decisions significantly affect individuals and ensure that individuals are given the opportunity and sufficient support to challenge or respond to such decisions.
- Consult with the public and civil society organizations regarding the appropriateness and impact of any proposed use of profiling and provide them with an opportunity to comment on the effects the profiling may have on society and individuals’ lives.

APPENDIX A: RESOURCES

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada. *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans,* December 2014. http://www.pre.ethics.gc.ca/pdf/eng/tcps2-2014/TCPS_2_FINAL_Web.pdf.

Crawford, Kate. “The Hidden Biases in Big Data,” *Harvard Business Review*, April 2013. https://hbr.org/2013/04/the-hidden-biases-in-big-data.

Dwork, Cynthia and Deirdre K. Mulligan. “It’s Not Privacy, and It’s Not Fair,” *Standford Law Review Online* 66, no. 35 (2013). http://scholarship.law.berkeley.edu/facpubs/2622.

European Union. *General Data Protection Regulation*, 2016. http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf.

Goodman, Bryce and Seth Flaxman. “European Union regulations on algorithmic decision-making and a ‘right to explanation’,” August 2016. https://arxiv.org/pdf/1606.08813.pdf.

Metcalf, Jacob, Emily F. Keller and danah boyd. *Perspectives on Big Data, Ethics, and Society*, May 2016. http://bdes.datasociety.net/wp-content/uploads/2016/05/Perspectives-on-Big-Data.pdf.

Ontario, Canada. Office of the Information and Privacy Commissioner. *De-identification Guidelines for Structured Data*, June 2016. https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf.

Ontario, Canada. Commission on Freedom of Information and Individual Privacy (Williams Commission). *Public Government for Private People: The Report of the Commission on Freedom of Information and Individual Privacy*, vol. 3. Toronto: Queen’s Printer, 1980.

United Kingdom. Information Commissioner’s Office. *Big data, artificial intelligence, machine learning and data protection*, March 2017. https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

United States. Executive Office of the President. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, May 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

United States. Federal Trade Commission. *Big Data: A Tool for Inclusion or Exclusion?* January 2016. https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.

United States. Privacy Protection Study Commission. *Personal Privacy in an Information Society*, 1977. https://epic.org/privacy/ppsc1977report/.

Big Data Guidelines

Information and Privacy
Commissioner of Ontario

Commissaire à l'information et à la
protection de la vie privée de l'Ontario

2 Bloor Street East, Suite 1400
Toronto, Ontario
Canada M4W 1A8

www.ipc.on.ca
416-326-3333
info@ipc.on.ca

May 2017